# Introduction to Monte Carlo Statistical Methods

George Casella
University of Florida

Exerpts from the book

Monte Carlo Statistical Methods

by

Christian Robert and George Casella

Springer-Verlag 1999

# Contents

# CHAPTER 1

## Introduction

- Experimenters choice before fast computers
  - Describe an accurate model which would usually preclude the computation of explicit answers
  - or choose a standard model which would allow this computation, but may not be a close representation of a realistic model.
- Such problems contributed to the development of simulation-based inference

**1.1 Statistical Models**

# Example 1.1.1 –Censored data models–

— are missing data models where densities are not sampled directly.

In a typical simple statistical model, we would observe

$$Y_1, \cdots, Y_n \sim f(y|\theta).$$

The distribution of the sample would then be given by the product

$$\prod_{i=1}^{n} f(y_i|\theta).$$

Inference about $\theta$ would then be based on this distribution.

With *censored* random variables the actual observations are

$$Y_i^* = \min\{Y_i, \overline{u}\}$$

where $\overline{u}$ is censoring point.

As a particular example, if

$$X \sim \mathcal{N}(\theta, \sigma^2)\text{and}Y \sim \mathcal{N}(\mu, \rho^2),$$

the variable

$$Z = X \wedge Y = \min(X, Y)$$

is distributed as

$$\left[1-\Phi\left(\frac{z-\theta}{\sigma}\right)\right] \times \rho^{-1}\varphi\left(\frac{z-\mu}{\rho}\right) \\ +\left[1-\Phi\left(\frac{z-\mu}{\rho}\right)\right] \sigma^{-1}\varphi\left(\frac{z-\theta}{\sigma}\right)$$

where $\varphi$ and $\Phi$ are the density and cdf of the normal $\mathcal{N}(0,1)$ distribution.

Similarly, if

$$X \sim \text{Weibull}(\alpha,\beta),$$

with density

$$f(x) = \alpha\beta x^{\alpha-1}\exp(-\beta x^{\alpha})$$

the censored variable

$$Z = X \wedge \omega, \quad \omega \text{constant},$$

has the density

$$f(z) = \alpha\beta z^{\alpha} e^{-\beta z^{\alpha}}\ \mathbb{I}_{z\leq\omega} + \left(\int_{\omega}^{\infty}\ \alpha\beta x^{\alpha} e^{-\beta x^{\alpha}} dx\right) \delta_{\omega}(z)\ ,$$

where $\delta_a(\cdot)$ is the Dirac mass at $a$. $\|$

**Example 1.1.2 –Mixture models–**

Models of *mixtures of distributions* are based on the assumption

$$X \sim f_j \text{ with probability } p_j,$$

for $j = 1, 2, \ldots, k$, with overall density

$$X \sim p_1 f_1(x) + \cdots + p_k f_k(x)\ .$$

If we observe a sample of independent random variables $(X_1, \cdots, X_n)$, the sample density is

$$\prod_{i=1}^{n} \{p_1 f_1(x_i) + \cdots + p_k f_k(x_i)\}\ .$$

Expanding this product shows that it involves $k^n$ elementary terms, which is prohibitive to compute in large samples. ||

**Example 1.1.3 –Student's $t$ distribution–** An reasonable alternative to normal errors is the Student's $t$ distribution, denoted by $\mathcal{T}(p, \theta, \sigma)$, which is often more "robust" against possible modeling errors (and others). The density of $\mathcal{T}(p, \theta, \sigma)$ is proportional to

$$\sigma^{-1}\left(1+\frac{(x-\theta)^2}{p\sigma^2}\right)^{-(p+1)/2},$$

If $p$ is known and the parameters $\theta$ and $\sigma$ are unknown, the likelihood is

$$\sigma^{n\frac{p+1}{2}}\prod_{i=1}^{n}\left(1+\frac{(x_i-\theta)^2}{p\sigma^2}\right).$$

This polynomial of degree $2n$ may have $n$ local minima, each of which needs to be calculated to determine the global maximum.

Illustration of the multiplicity of modes of the likelihood from a Cauchy distribution $\mathcal{C}(\theta, 1)$ $(p = 1)$ when $n = 3$ and $X_1 = 0$, $X_2 = 5$, $X_3 = 9$. $\|$

5.5in5.5in/work/short/mcmcv22/figures/bmp/cauchy.bmp

Figure 1.1.1. *Likelihood of the sample* $(0, 5, 9)$ *from the distribution* $\mathcal{C}(\theta, 1)$.

CHAPTER 2

# Random Variable Generation

- We rely on the possibility of producing (with a computer) a supposedly endless flow of random variables (usually iid) for well-known distributions.
- We look at a uniform random number generator and illustrate methods for using these uniform random variables to produce random variables from both standard and nonstandard distributions

## 2.1 Basic Methods

### 2.1.1 *Desiderata and Limitations*

> "Any one who considers arithmetical methods of reproducing random digits is, of course, in a state of sin. As has been pointed out several times, there is no such thing as a random number—there are only methods of producing random numbers, and a strict arithmetic procedure of course is not such a method." –John Von Neumann (1951)

- The problem is to produce a *deterministic* sequence of values in $[0, 1]$ which imitates a sequence of *iid* uniform random variables $\mathcal{U}_{[0,1]}$.
- Can't use the physical imitation of a "random draw" (no guarantee of uniformity, no reproducibility)
- random sequence in the following sense: Having generated $(X_1, \cdots, X_n)$, knowledge of $X_n$ [or of $(X_1, \cdots, X_n)$] imparts no discernible knowledge of the value of $X_{n+1}$.
- Of course, given the initial value $X_0$, the sample $(X_1, \cdots, X_n)$ is always the same.
- the validity of a random number generator is based on a single sample $X_1, \cdots, X_n$ when $n$ tends to $+\infty$ and not on replications $(X_{11}, \cdots, X_{1n})$, $(X_{21}, \cdots, X_{2n}), \ldots (X_{k1}, \cdots, X_{kn})$ where $n$ is fixed and $k$ tends to infinity.
- In fact, the distribution of these $n$-tuples depends on the manner in which the initial values $X_{r1}$ $(1 \le r \le k)$ were generated.

## 2.2 Transformation Methods

- The case where a distribution $f$ is linked in a relatively simple way to another distribution that is easy to simulate.

**Example 2.2.1 –Exponential variables–**
If $U \sim \mathcal{U}_{[0,1]}$, the random variable

$$X = -\log U/\lambda$$

has distribution

$$\begin{aligned} P(X \le x) &= P(-\log U \le \lambda x) \\ &= P(U \ge e^{-\lambda x}) \\ &= 1 - e^{-\lambda x}, \end{aligned}$$

the exponential distribution $\mathcal{E}xp(\lambda)$. ||

- Other random variables that can be generated starting from an exponential include
  - $$Y = -2 \sum_{j=1}^{\nu} \log(U_j) \sim \chi^2_{2\nu}$$
  - $$Y = -\beta \sum_{j=1}^{a} \log(U_j) \sim \mathcal{G}a(a, \beta)$$
  - $$Y = \frac{\Sigma_{j=1}^{a} \log(U_j)}{\Sigma_{j=1}^{a+b} \log(U_j)} \sim \mathcal{B}e(a, b)$$

## 2.3 Accept-Reject Methods

- There are many distributions from which it is difficult, or even impossible, to **directly** simulate.

- We now turn to another class of methods that only requires us to know the functional form of the density $f$ of interest up to a multiplicative constant.

- The key to this method is to use a simpler (simulation-wise) density $g$ from which the simulation is actually done.

  - For a given density $g$
    — the *instrumental density*
  - there are many densities $f$
    —the *target densities*

  which can be simulated this way.

- We first look at the *Accept-Reject method.*
  - Given a density of interest $f$,
  - find a density $g$ and a constant $M$ such that

$$f(x) \leq Mg(x)$$

on the support of $f$.

  - **Algorithm A.1 –Accept-Reject Method–**

1. `Generate` $X \sim g$, $U \sim \mathcal{U}_{[0,1]}$ ;
2. `Accept` $Y = X$ `if` $U \leq f(X)/Mg(X)$ ;
3. `Return to 1. otherwise.`

*This produces a variable $Y$ distributed according to $f$.*

- This Algorithm has two interesting properties.
  - First, it provides a generic method to simulate from any density $f$ that is known *up to a multiplicative factor*.
    - This property is particularly important in Bayesian calculations. There the posterior distribution is
      $$\pi(\theta|x) \propto \pi(\theta)\ f(x|\theta)\ .$$
      which is easily specified up to a normalizing constant
  - A second property of the lemma is that the probability of acceptance in the algorithm is exactly $1/M$.
    - The expected number of trials until a variable is accepted is $M$

## Example 2.3.1 –Normal from a Cauchy–

- $$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

  and

  $$g(x) = \frac{1}{\pi} \frac{1}{1+x^2},$$

  densities of the normal and Cauchy distributions.

- $$\frac{f(x)}{g(x)} = \sqrt{\frac{\pi}{2}}(1+x^2)\, e^{-x^2/2} \le \sqrt{\frac{2\pi}{e}} = 1.52$$

  attained at $x = \pm 1$.

- So the probability of acceptance $1/1.52 = 0.66$, and, on the average, one out of every three simulated Cauchy variables is rejected.

- The mean number of trials to success is 1.52.

$\|$

## Example 2.3.2 Gamma with non-integer shape parameter

- This illustrates a real advantage of the Accept-Reject algorithm.
- the gamma distribution $\mathcal{G}a(\alpha, \beta)$ can be represented as the sum of $\alpha$ exponential random variables.
- This is impossible if $\alpha$ is not an integer
- Can use the Accept-Reject algorithm with instrumental distribution

  $$\mathcal{G}a(a, b), \text{ with } a = [\alpha], \quad \alpha \geq 0.$$

  (Without loss of generality, $\beta = 1$.)
- Up to a normalizing constant,

  $$f/g_b = b^{-a} x^{\alpha - a} \, \exp\{-(1-b)x\} \leq b^{-a} \left( \frac{\alpha - a}{(1-b)e} \right)^{\alpha - a}$$

  for $b \leq 1$.
  The maximum is attained at $b = a/\alpha$.

$\|$

## Example 2.3.3 Truncated Normal distributions.

- *Truncated Normals* appear in many contexts

- When constraints $x \geq \underline{\mu}$ produce densities proportional to

  $$e^{-(x-\mu)^2/2\sigma^2}\ \mathbb{I}_{x \geq \underline{\mu}}$$

  for a bound $\underline{\mu}$ large compared with $\mu$,

  - there are alternatives which are far superior to the naïve method of generating a $\mathcal{N}(\mu, \sigma^2)$ until exceeding $\underline{\mu}$.
  - This approach requires an average number of $1/\Phi((\mu - \underline{\mu})/\sigma)$ simulations from $\mathcal{N}(\mu, \sigma^2)$ for one acceptance.

- An instrumental distribution is the translated exponential distribution, $\mathcal{E}xp(\alpha, \underline{\mu})$, with density

  $$g_\alpha(z) = \alpha e^{-\alpha(z-\underline{\mu})}\ \mathbb{I}_{z \geq \underline{\mu}}\ .$$

- The ratio $f/g_\alpha$ is then bounded by

  $$f/g_\alpha \leq \begin{cases} 1/\alpha\ \exp(\alpha^2/2 - \alpha\underline{\mu}) & \text{if } \alpha > \underline{\mu}, \\ 1/\alpha\ \exp(-\underline{\mu}^2/2) & \text{otherwise.} \end{cases}$$

$\|$

# CHAPTER 3

## Monte Carlo Integration

- Two major classes of numerical problems that arise in statistical inference
  - optimization - generally associated with the likelihood approach
  - integration- generally associated with the Bayesian approach

### 3.1 Importance Sampling

- Simulation from $f$ (the true density) is not necessarily optimal, in fact, it is usually sub-optimal.
- The alternative to direct sampling from $f$ is *importance sampling*.

**Definition 3.1.1** The method of *importance sampling* is an evaluation of

$$\mathbb{E}_f[h(X)] = \int_{\mathcal{X}} h(x)\, f(x)\, dx \ .$$

based on generating a sample $X_1, \ldots, X_n$ from a given distribution $g$, and approximating

$$\mathbb{E}_f[h(X)] \approx \frac{1}{m} \sum_{j=1}^{m} \frac{f(X_j)}{g(X_j)}\, h(X_j) \ .$$

This method is based on the alternative representation

$$\mathbb{E}_f[h(X)] = \int_{\mathcal{X}} \left[ h(x)\, \frac{f(x)}{g(x)} \right] g(x)\, dx \ .$$

- The estimator

$$\mathbb{E}_f[h(X)] \approx \frac{1}{m} \sum_{j=1}^{m} \frac{f(X_j)}{g(X_j)}\, h(X_j)$$
$$\to \int_{\mathcal{X}} h(x)\, f(x)\, dx$$

  - converges for same reason the regular Monte Carlo estimator $\overline{h}_m$ converges;
  - converges for any choice of the distribution $g$ [as long as $\mathrm{supp}(g) \supset \mathrm{supp}(f)$].
  - The instrumental distribution $g$ can be chosen from distributions that are easy to simulate.
  - The same sample (generated from $g$) can be used repeatedly, not only for different functions $h$ but also for different densities $f$.

**Example 3.1.2 –Student's $t$ distribution–**
Consider $X \sim \mathcal{T}(\nu, \theta, \sigma^2)$, with density

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\sigma\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{(x-\theta)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2} .$$

Without loss of generality, take $\theta = 0$, $\sigma = 1$.

- Calculate the integral
$$\int_{2.1}^{\infty} x^5 f(x) dx.$$

- Simulation possibilities
  - Directly from $f$, since $f = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_\nu}}$
  - Importance sampling using Cauchy $\mathcal{C}(0, 1)$
  - Importance sampling using a normal (expected to be nonoptimal).
  - Importance sampling using a $\mathcal{U}([0, 1/2.1])$

- The figure shows
  - Uniform is best
  - Cauchy is OK
  - $f$ and Normal are rotten

$\|$

# CHAPTER 4

## Markov Chains

- Use of Markov chains
  - Many algorithms can be described as Markov chains
- Needed properties
  - The quantity of interest is what the chain converges to
- We need to know
  - When will chains converge
  - What do they converge to

### 4.1 Basic notions

- A *Markov chain* is a sequence of random variables that can be thought of as evolving over time.
- The probability of a transition depending on the particular set that the chain is in
- We define the chain in terms of its *transition kernel*, the function that determines these transitions.

**Definition 4.1.1** A *transition kernel* is a function $K$ defined on $\mathcal{X} \times \mathcal{B}(\mathcal{X})$ such that

(i) $\forall x \in \mathcal{X}$, $K(x, \cdot)$ is a probability measure;

(ii) $\forall A \in \mathcal{B}(\mathcal{X})$, $K(\cdot, A)$ is measurable.

- When $\mathcal{X}$ is *discrete*, the transition kernel simply is a (transition) matrix $K$ with elements
$$P_{xy} = P(X_n = y | X_{n-1} = x) \ , \qquad x, y \in \mathcal{X}.$$
- In the continuous case, the *kernel* also denotes the conditional density $K(x, x')$ of the transition $K(x, \cdot)$. That is,
$$P(X \in A | x) = \int_A K(x, x') dx'.$$

**Definition 4.1.2** Given a transition kernel $K$, a sequence $X_0, X_1, \ldots, X_n, \ldots$ of random variables is a *Markov chain*, denoted by $(X_n)$, if, for any $t$, the conditional distribution of $X_t$ given $x_{t-1}, x_{t-2}, \ldots, x_0$ is the same as the distribution of $X_t$ given $x_{t-1}$. That is,

$$\begin{aligned} P(X_{k+1} \in A | x_0, x_1, x_2, \ldots, x_k) & \\ &= P(X_{k+1} \in A | x_k) \\ &= \int_A K(x_k, dx) \end{aligned}$$

### 4.2 Ergodicity and convergence

- We consider: *to what is the chain converging?*

- The invariant distribution $\pi$ is the natural candidate for the *limiting distribution*

- A fundamental property is *ergodicity*, or independence of initial conditions.

  - In the discrete case with a state $\omega$ is *ergodic* if
$$\lim_{n\to\infty} |K^n(\omega,\omega) - \pi(\omega)| = 0 \ .$$

- In general , we establish convergence using the *total variation norm*,

$$\|\mu_1 - \mu_2\|_{TV} = \sup_A \ |\mu_1(A) - \mu_2(A)|.$$

- and we want
$$\| \int \ K^n(x,\cdot)\mu(dx) - \pi \ \|_{TV}$$
$$= \sup_A | \int \ K^n(x,A)\mu(dx) - \pi(A) \ |$$
  to be small.

**Theorem 4.2.1** *If $(X_n)$ is Harris positive recurrent and aperiodic, then*

$$\lim_{n\to\infty} \|\int K^n(x,\cdot)\mu(dx) - \pi\|_{TV} = 0$$

*for every initial distribution $\mu$.*

- We thus take "*Harris positive recurrent and aperiodic*" as equivalent to "*ergodic*"
- Convergence in total variation implies
  $$\lim_{n\to\infty} |\mathbb{E}_\mu[h(X_n)] - \mathbb{E}^\pi[h(X)]| = 0$$
  for every *bounded function h*.
- There are difference speeds of convergence
  - ergodic (fast)
  - *geometrically* ergodic (faster)
  - *uniformly* ergodic (fastest)

## 4.3 Limit theorems

- Ergodicity determines the probabilistic properties of *average* behavior of the chain.

- But we also want to do *statistical inference*, which must reason by induction from the observed sample.

- The fact that $\|P_x^n - \pi\|$ is close to 0 does not bring direct information about
$$X_n \sim P_x^n$$
.

- We need LLNs and CLTs

- The classical LLNs and CLTs are not directly applicable due to:

  - The Markovian dependence structure between the observations $X_i$

  - The non-stationarity of the sequence.

**Theorem 4.3.1 Ergodic Theorem –LLN**

*If the Markov chain* $(X_n)$ *is Harris recurrent, then for any function* $h$ *with* $\mathrm{E}|h| < \infty$,

$$\lim_{n\to\infty} \frac{1}{n} h(X_i) = \int h(x) d\pi(x),$$

- To get a CLT, we need more assumptions.
- For MCMC, the easiest is *reversibility*

**Definition 4.3.2** A Markov chain $(X_n)$ is *reversible* if for all $n$

$$X_{n+1}|X_{n+2} \sim X_{n+1}|X_n.$$

- So the direction of time does not matter.

**Theorem 4.3.3** *If the Markov chain* $(X_n)$ *is Harris recurrent and reversible,*

$$\frac{1}{\sqrt{N}} \left( \sum_{n=1}^{N} \left( h(X_n) - \mathbb{E}^{\pi}[h] \right) \right) \overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0, \gamma_h^2) \ .$$

*where*

$$\begin{aligned} 0 < \gamma_h^2 \ &= \ \mathbb{E}_\pi[\overline{h}^2(X_0)] \\ &\quad +2 \ \sum_{k=1}^{\infty} \ \mathbb{E}_\pi[\overline{h}(X_0)\overline{h}(X_k)] < +\infty. \end{aligned}$$

# CHAPTER 5

## Monte Carlo Optimization

### 5.1 Introduction

- Differences between the *numerical approach* and the *simulation approach* to the problem
$$\max_{\theta \in \Theta} h(\theta)$$
lie in the treatment of the function $h$.

- Using *deterministic numerical methods*, the analytical properties of the target function (convexity, boundedness, smoothness) are often paramount.

- For the *simulation approach*, we are more concerned with $h$ from a probabilistic (rather than analytical) point of view.

**Example 5.1.1 Minimization.**

Consider minimizing

$$\begin{aligned} h(x,y) &= (x\sin(20y)+y\sin(20x))^2\cosh(\sin(10x)x) \\ &+ (x\cos(10y)-y\sin(10x))^2\cosh(\cos(20y)y)\ , \end{aligned}$$

with global minimum 0 at $(x,y)=(0,0)$.

- Many local minima.
- Standard methods may not find the global minimum
- We can simulate from $\exp(-h(x,y))$.
- Get the minimum from the resulting $h(x_i,y_i)$'s.

$\|$

- Use the stochastic gradient method with our test function
- Results of three stochastic gradient runs for the minimization of the function $h$ in Example 5.1.1 with different values of $(\alpha_j,\beta_j)$ and starting point $(0.65, 0.8)$. The iteration $T$ is obtained by the stopping rule $||\theta_T-\theta_{T-1}||<10^{-5}$.

5in4in/work/short/mcmcv22/figures/bmp/grid$_{m}ax.bmp$

Figure 5.1.1. *Grid representation of the function* $h(x, y)$ *of Example 5.1.1 on* $[-1, 1]^2$.

| | | | |
|---|---|---|---|
| $\alpha_j$ | $1/10j$ | $1/100j$ | $1/10\log(1+j)$ |
| $\beta_j$ | $1/10j$ | $1/100j$ | $1/j$ |
| $\theta_T$ | $(-0.166, 1.02)$ | $(0.629, 0.786)$ | $(0.0004, 0.245)$ |
| $h(\theta_T)$ | $1.287$ | $0.00013$ | $4.24 \times 10^{-6}$ |
| $\min_t h(\theta_t)$ | $0.115$ | $0.00013$ | $2.163 \times 10^{-7}$ |
| Iteration | 50 | 93 | 58 |

- **Simulated Annealing**
  - This name is borrowed from Metallurgy: A metal manufactured by a slow decrease of temperature (*annealing*) is stronger than a metal manufactured by a fast decrease of temperature.
- Fundamental idea: A change of scale, called *temperature*, allows greater exploration $h$
- Rescaling partially avoids trapping in local maxima.
- Given a temperature $T > 0$, generate

$$\theta_1^T, \theta_2^T \sim \pi(\theta) \propto \exp(h(\theta)/T)$$

  and approximate the maximum of $h$.

  - As $T \downarrow 0$, the values simulated concentrate in a narrower and narrower neighborhood of the local maxima of $h$

- The **Algorithm** proposed by Metropolis *et al.* (1953).

- Starting from $\theta_0$,

  ∘ $\zeta \sim$ uniform in a neighborhood of $\theta_0$

  ∘ the new value of $\theta$ is generated by:

  $$\theta_1 = \begin{cases} \zeta & \text{with probability } \rho = \exp(\Delta h/T) \wedge 1 \\ \theta_0 & \text{with probability } 1 - \rho, \end{cases}$$

  where $\Delta h = h(\zeta) - h(\theta_0)$. .

- Therefore,

  ∘ if $h(\zeta) \geq h(\theta_0)$, $\zeta$ is accepted with probability 1

  ∘ if $h(\zeta) < h(\theta_0)$, $\zeta$ may still be accepted with probability $\rho \neq 0$

- So if $\theta_0$ is a local maximum of $h$, the algorithm escapes with a probability that depends on $T$

- Usually, the simulated annealing algorithm modifies the temperature $T$ at each iteration.

- **The EM Algorithm**
- introduced by Dempster *et al.* (1977) to overcome the difficulties in maximizing likelihoods
- taking advantage of the representation

$$g(x|\theta) = \int_{\mathcal{Z}} f(x, z|\theta)\, dz$$

  and solving a sequence of easier maximization problems whose limit is the answer to the original problem.
- EM algorithm relates to MCMC algorithms in the sense that it can be seen as a forerunner of the Gibbs sampler in its Data Augmentation version, replacing simulation by maximization.
- Suppose that we observe $X_1, \ldots, X_n$, iid from $g(x|\theta)$ and want to compute

$$\hat{\theta} = \arg\max L(\theta|\mathbf{x}) = \prod_{i=1}^{n} g(x_i|\theta).$$

- We augment the data with $\mathbf{z}$, where $\mathbf{X}, \mathbf{Z} \sim f(\mathbf{x}, \mathbf{z}|\theta)$ and note the identity

$$k(\mathbf{z}|\theta, \mathbf{x}) = \frac{f(\mathbf{x}, \mathbf{z}|\theta)}{g(\mathbf{x}|\theta)},$$

  where $k(\mathbf{z}|\theta, \mathbf{x})$ is the conditional distribution of the missing data $\mathbf{Z}$ given the observed data $\mathbf{x}$.

- This identity leads to the following relationship between the complete-data likelihood

$$L^c(\theta|\mathbf{xz}) = f(\mathbf{x}, \mathbf{z}|\theta)$$

and the observed data likelihood

$$L(\theta|\mathbf{x}).$$

For any value $\theta_0$,

$$\begin{aligned}\log L(\theta|\mathbf{x}) \;&=\; \mathbb{E}_{\theta_0}[\log L^c(\theta|\mathbf{x}, \mathbf{z})|\theta_0, \mathbf{x}] \\ &\quad -\mathbb{E}_{\theta_0}[\log k(\mathbf{z}|\theta, \mathbf{x})|\theta_0, \mathbf{x}],\end{aligned}$$

where the expectation is with respect to $k(\mathbf{z}|\theta_0, \mathbf{x})$.

- the strength of the EM algorithm is that we only have to deal with the first term on the right side above, as the other term can be ignored.

- The likelihood is increased at every iteration - there are convergence guarantees

- Denote the expected log-likelihood by
$$Q(\theta|\theta_0, \mathbf{x}) = \mathbb{E}_{\theta_0}[\log L^c(\theta|\mathbf{x}, \mathbf{z})|\theta_0, \mathbf{x}].$$

- a sequence of estimators $\hat{\theta}_{(j)}$, $j = 1, 2, \ldots$, is obtained iteratively by
$$Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j-1)}, \mathbf{x}) = \max_{\theta} Q(\theta|\hat{\theta}_{(j-1)}, \mathbf{x}).$$

**Algorithm A.2 –The EM Algorithm–**

1. (*the E-step*) `Compute`
$$Q(\theta|\hat{\theta}_{(m)}, \mathbf{x}) = \mathbb{E}_{\hat{\theta}_{(m)}}[\log L^c(\theta|\mathbf{x}, \mathbf{z})] \,,$$
`where the expectation is with respect to` $k(\mathbf{z}|\hat{\theta}_m, \mathbf{x})$ .

2. (*the M-step*) `Maximize` $Q(\theta|\hat{\theta}_{(m)}, \mathbf{x})$ `in` $\theta$ `and take`
$$\theta_{(m+1)} = \arg\max_{\theta} \ Q(\theta|\hat{\theta}_{(m)}, \mathbf{x}).$$

The iterations are conducted until a fixed point of $Q$ is obtained.

## Example 5.1.2 Censored data

If $f(x-\theta)$ is the $\mathcal{N}(\theta,1)$ density, the censored data likelihood is

$$L(\theta|\mathbf{x}) = \frac{1}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta)^2\right\}[1-\Phi(a-\theta)]^{n-m}$$

and the complete-data log-likelihood is

$$\log L^c(\theta|\mathbf{x},\mathbf{z}) \propto -\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta)^2 - \frac{1}{2}\sum_{i=m+1}^{n}(z_i-\theta)^2 \ ,$$

where the $z_i$'s are observations from the truncated Normal distribution

$$k(z|\theta,\mathbf{x}) = \frac{\exp\{-\frac{1}{2}(z-\theta)^2\}}{\sqrt{2\pi}[1-\Phi(a-\theta)]} = \frac{\varphi(z-\theta)}{1-\Phi(a-\theta)}, \qquad a < z.$$

At the $j$th step in the EM sequence, we have

$$\begin{aligned} Q(\theta|\hat{\theta}_{(j)},\mathbf{x}) \ \propto \ & -\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta)^2 \\ & -\frac{1}{2}\sum_{i=m+1}^{n}\int_a^\infty (z_i-\theta)^2 k(z|\hat{\theta}_{(j)},\mathbf{x})\,dz_i, \end{aligned}$$

Differentiating with respect to $\theta$ yields

$$\hat{\theta}_{(j+1)} = \frac{m\bar{x} + (n-m)\mathbb{E}[Z|\hat{\theta}_{(j)}]}{n} \; ,$$

where

$$\mathbb{E}[Z|\hat{\theta}_{(j)}] = \int_a^\infty zk(z|\hat{\theta}_{(j)}, \mathbf{x})\, dz = \hat{\theta}_{(j)} + \frac{\varphi(a - \hat{\theta}_{(j)})}{1 - \Phi(a - \hat{\theta}_{(j)})}.$$

Thus, the EM sequence is defined by

$$\hat{\theta}_{(j+1)} = \frac{m}{n}\bar{x} + \frac{n-m}{n}\left[\hat{\theta}_{(j)} + \frac{\varphi(a - \hat{\theta}_{(j)})}{1 - \Phi(a - \hat{\theta}_{(j)})}\right],$$

which converges to the MLE $\hat{\theta}$. ||

- A (sometime) difficulty with the EM algorithm is the computation of $Q(\theta|\theta_0, \mathbf{x})$.
- To overcome this difficulty, use
$$\hat{Q}(\theta|\theta_0, \mathbf{x}) = \frac{1}{m}\sum_{i=1}^m \log L^c(\theta|\mathbf{x}, \mathbf{z}) \; ,$$
where $Z_1, \ldots, Z_m \sim k(\mathbf{z}|\mathbf{x}, \theta)$.
- When $m \to \infty$, this quantity converges to $Q(\theta|\theta_0, \mathbf{x})$.

# CHAPTER 6

## The Metropolis-Hastings Algorithm

### 6.1 Monte Carlo Methods based on Markov Chains

- We know it is not necessary to use a sample from the distribution $f$ to approximate the integral
$$\int h(x)f(x)dx \ ,$$
- Now we obtain $X_1, \ldots, X_n \sim f$ **(approx)** without directly simulating from $f$.
  - We use an *ergodic Markov chain* with stationary distribution $f$

- For an arbitrary starting value $x^{(0)}$, an ergodic chain $(X^{(t)})$ is generated using a transition kernel with stationary distribution $f$
- This insures the convergence in distribution of $(X^{(t)})$ to a random variable from $f$.
- For a "large enough" $T_0$, $X^{(T_0)}$ can be considered as distributed from $f$
- We thus produce a *dependent* sample $X^{(T_0)}, X^{(T_0+1)}, \ldots,$ which is generated from $f$.

### 6.2 The Metropolis–Hastings algorithm

- The algorithm starts with the objective (target) density $f$
- A conditional density $q(y|x)$, called the *instrumental* (or *proposal*) *distribution*, is then chosen.
- **Algorithm A.3 –Metropolis–Hastings–**

`Given` $x^{(t)}$,

1. `Generate` $Y_t \sim q(y|x^{(t)})$.
2. `Take`

$$X^{(t+1)} = \begin{cases} Y_t & \texttt{with prob.} \quad \rho(x^{(t)}, Y_t), \\ x^{(t)} & \texttt{with prob.} \quad 1 - \rho(x^{(t)}, Y_t), \end{cases}$$

`where`

$$\rho(x, y) = \min\left\{ \frac{f(y)}{f(x)} \, \frac{q(x|y)}{q(y|x)} \, , 1 \right\} .$$

# Example 6.2.1 –Saddlepoint tail area approximation–

- Saddlepoint approximation are useful for non-central chi squared tail areas.

- An alternative is to sample $Z_1, \ldots, Z_m$, from the saddlepoint distribution, and use

$$P(\bar{X} > a)$$
$$= \int_{\hat{\tau}(a)}^{\infty} \left(\frac{n}{2\pi}\right)^{1/2} \left[K_X''(t)\right]^{1/2} \exp\left\{n\left[K_X(t) - tK_X'(t)\right]\right\} dt$$
$$\approx \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}[Z_i > \hat{\tau}(a)] \, ,$$

  - where $K_X(\tau)$ is the cumulant generating function of $X$

  - $\hat{\tau}(x)$ is the solution of $K'(\hat{\tau}(x)) = x$ .

- We can derive an instrumental density to use in a Metropolis–Hastings algorithm. Using a Taylor series approximation,

$$\exp\left\{n\left[K_X(t) - tK_X'(t)\right]\right\} \approx \exp\left\{-nK_X''(0)\frac{t^2}{2}\right\}$$

  - a first choice for an instrumental density is the $\mathcal{N}(0, 1/nK_X''(0))$

- Use M-H with normal candidate density and
$$K''_X(t) = 2[p(1-2t) + 4\lambda]/(1-2t)^3.$$
  - The same set of simulated random variables are used for all calculations.
  - We avoid calculating the saddlepoint normalizing constant

- Monte Carlo saddlepoint approximation of a noncentral chi squared integral for $p = 6$ and $\lambda = 9$, based on $10,000$ simulated random variables.

| interval | renormalized saddlepoint | exact | Monte Carlo |
|---|---|---|---|
| $(36.225, \infty)$ | .0996 | .1 | .0992 |
| $(40.542, \infty)$ | .0497 | .05 | .0497 |
| $(49.333, \infty)$ | .0099 | .01 | .0098 |

‖

- **There are many other algorithms**
  - *Adaptive Rejection Metropolis Sampling*
  - *Reversible Jumps*
  - *Langevin algorithms*
  - to name a few...

# CHAPTER 7

## The Gibbs Sampler

### 7.1 General Principles

- A very specific simulation algorithm based on the target $f$.
- Uses the conditional densities $f_1, \ldots, f_p$ from $f$
- Start with the random variable $\mathbf{X} = (X_1, \ldots, X_p)$
- Simulate from the conditional densities,

$$X_i | x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p$$
$$\sim f_i(x_i | x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$$

for $i = 1, 2, \ldots, p$.

## • Algorithm A.4 –The Gibbs sampler–

`Given` $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$, `generate`

1. $X_1^{(t+1)} \sim f_1(x_1|x_2^{(t)}, \ldots, x_p^{(t)})$;

2. $X_2^{(t+1)} \sim f_2(x_2|x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)})$,

$\ldots$

p. $X_p^{(t+1)} \sim f_p(x_p|x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)})$,

*then* $\mathbf{X}^{(t+1)} \to \mathbf{X} \sim f$.

- ○ The densities $f_1, \ldots, f_p$ are called the *full conditionals*
- ○ these are the only densities used for simulation
- ○ Thus, even in a high dimensional problem, *all of the simulations may be univariate*

**Example 7.1.1 –Cauchy-normal –**
Consider the density

$$f(\theta|\theta_0) \propto \frac{e^{-\theta^2/2}}{[1+(\theta-\theta_0)^2]^\nu}.$$

This is the posterior distribution resulting from the model

$$X|\theta \sim \mathcal{N}(\theta,1) \text{ and } \theta \sim \mathcal{C}(\theta_0,1).$$

The density $f(\theta|\theta_0)$ can be written as the marginal density

$$f(\theta|\theta_0) \propto \int_0^\infty \ e^{-\theta^2/2}\ e^{-[1+(\theta-\theta_0)^2]\ \eta/2}\ \eta^{\nu-1}\ d\eta,$$

and can therefore be completed as

$$g(\theta,\eta) \propto e^{-\theta^2/2}\ e^{-[1+(\theta-\theta_0)^2]\ \eta/2}\ \eta^{\nu-1},$$

which leads to the conditional densities

$$\begin{aligned} g_1(\eta|\theta) &= \mathcal{G}a\left(\nu, \frac{1+(\theta-\theta_0)^2}{2}\right), \\ g_2(\theta|\eta) &= \mathcal{N}\left(\frac{\theta_0\eta}{1+\eta}, \frac{1}{1+\eta}\right). \end{aligned}$$

Note that the parameter $\eta$ is completely meaningless for the problem at hand but serves to facilitate computations. ) $\|$

- The Gibbs sampler is particularly well suited to *hierarchical models.*
- Such models naturally appear in Bayesian analysis

**Example 7.1.2 –Hierarchical models in animal epidemiology–**

- Schukken *et al.* (1991) obtained counts of the number of cases of clinical mastitis in 127 dairy cattle herds over a one year period.
  - $X_i$, $i = 1, \cdots, m$, denote the number of cases in herd $i$
  - $X_i \sim \mathcal{P}(\lambda_i)$, where $\lambda_i$ is the underlying rate of infection in herd $i$
  - Lack of independence here (mastitis is infectious) might manifest itself as overdispersion.
  - To account for this, they used the model

$$\begin{aligned} X_i &\sim \mathcal{P}(\lambda_i) \\ \lambda_i &\sim \mathcal{G}a(\alpha, \beta_i) \\ \beta_i &\sim \mathcal{IG}(a, b), \end{aligned}$$

  - The Gibbs sampler

$$\lambda_i \sim \pi(\lambda_i | \mathbf{x}, \alpha, \beta_i) = \mathcal{G}a(x_i + \alpha, [1 + 1/\beta_i]^{-1})$$
$$\beta_i \sim \pi(\beta_i | \mathbf{x}, \alpha, a, b, \lambda_i) = \mathcal{IG}(\alpha + a, [\lambda_i + 1/b]^{-1})$$

gives the posterior density of $\lambda_i$, $\pi(\lambda_i | \mathbf{x}, \alpha)$

$\|$

# CHAPTER 8

## Diagnosing Convergence

### 8.1 Stopping the Chain

- Convergence results do not tell us when to stop the MCMC algorithm and produce our estimates.
- We now look at methods of controlling the chain in the sense of a *stopping rule* to guarantee that the number of iterations is sufficient.
- From a general point of view, there are three (increasingly stringent) types of convergence for which assessment is necessary.

- *Convergence to the Stationary Distribution*
  - ◇ a minimal requirement for an algorithm that approximates simulation from $f$

- *Convergence of Averages* Here we are concerned with convergence of the empirical average

$$\frac{1}{T} \sum_{t=1}^{T} h(\theta^{(t)}) \longrightarrow \mathbb{E}_f[h(\theta)].$$

  - ◇ This type of convergence is most relevant in the implementation of MCMC algorithms.

- *Convergence to iid Sampling*
  - ◇ This measures how close a sample $(\theta_1^{(t)}, \ldots, \theta_n^{(t)})$ is to being iid.
  - ◇ the goal is to produce variables $\theta_i$ which are (quasi-)independent.

### 8.2 Monitoring Convergence to the Stationary Distribution

- **Graphical Methods**
- A natural empirical approach to convergence control is to draw pictures of the output of simulated chains

- This may detect deviant or nonstationary behaviors

- A first idea is to draw the sequence of the $\theta^{(t)}$'s against $t$

- However, this plot is only useful for strong nonstationarities of the chain.

**Example 8.2.1 –Witch's hat distribution–**
Consider

$$\pi(\theta|y) \propto \left\{(1-\delta)\,\sigma^{-d} e^{-\|y-\theta\|^2/(2\sigma^2)} + \delta\right\} \mathbb{I}_C(\theta), \quad y \in \mathbb{R}^d$$

when $\theta$ is in to the unit cube $C = [0,1]^d$.

- This density has a mode which is very concentrated around $y$ for small $\delta$ and $\sigma$

- The strong attraction of the mode gives the impression of stationarity for the chain
- The chain with initial value 0.9098, which achieves a momentary escape from the mode, is actually atypical.
- This example has become a *benchmark* to evaluate the performances of different methods of convergence. control.

$\|$

### 8.3 Monitoring Convergence of Averages

• *Multiple Estimates*

**Example 8.3.1 – Cauchy posterior** – For the posterior distribution

$$\pi(\theta|x_1,x_2,x_3) \;\propto\; e^{-\theta^2/2\sigma^2} \prod_{i=1}^{3} \frac{1}{1+(\theta-x_i)^2}.$$

a completion Gibbs sampling algorithm can be derived by introducing three artificial variables, $\eta_1,\eta_2,\eta_3$, such that

$$\pi(\theta,\eta_1,\eta_2,\eta_3|x_1,x_2,x_3) \;\propto\; e^{-\theta^2/2\sigma^2} \prod_{i=1}^{3} e^{-(1+(\theta-x_i)^2)\eta_i/2},$$

resulting in the Gibbs sampler ($i=1,2,3$)

$$\eta_i|\theta,x_i \;\sim\; \mathcal{E}xp\left(\frac{1+(\theta-x_i)^2}{2}\right),$$

$$\theta|x_1,x_2,x_3,\eta_1,\eta_2,\eta_3 \;\sim\; \mathcal{N}\left(\frac{\Sigma_i\,\eta_i x_i}{\Sigma_i\,\eta_i+\sigma^{-2}}\;,\;\frac{1}{\Sigma_i\,\eta_i+\sigma^{-2}}\right).$$

- The figure illustrates the efficiency of this algorithm by exhibiting the agreement between the histogram of the simulated $\theta^{(t)}$'s and the true posterior distribution
- If the function of interest is $h(\theta) = \exp(-\theta/\sigma)$, the different approximations of $\mathbb{E}_\pi[h(\theta)]$ can be monitored.

- The figure graphs the convergence of four estimators versus $T$ (plus one more).
- The strong agreement of $S_T$, $S_T^C$ indicates convergence
- The bad behavior the importance sampler is most likely associated with an infinite variance.

$\|$

# CHAPTER 9

## Implementation in Missing Data Models

### 9.1 Introduction

- Missing data models are a natural application for simulation
- Simulation replaces the missing data part so that one can proceed with a "classical" inference on the complete model.
- The EM algorithm that Dempster *et al.* (1977) first described a rigorous and general formulation of statistical inference though completion of missing data.
- Now we illustrate the potential of Markov Chain Monte Carlo algorithms in the analysis of missing data models

## Example 9.1.1 – Probit Regression –

- Another situation where grouped data appears in a natural fashion is that of *qualitative models*.

- We look at the probit model, often considered as a threshold model.

- We observe $Y_i \sim$ Bernoulli$\{0, 1\}$ and link them to a vector of covariates $x_i$ by the equation
$$p_i = \Phi(x_i^t \beta) \ , \qquad \beta \in \mathbb{R}^p.$$
where $\Phi$ is the standard normal cdf.

- The $Y_i$'s can be thought of as delimiting a threshold.

    - Assume there are latent (unobservable) continuous random variables $Y_i^*$ where
$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0, \\ 0 & \text{otherwise.} \end{cases}$$

    - Thus, $p_i = P(Y_i = 1) = P(Y_i^* > 0)$, and we have an automatic way to complete the model $\rightarrow$

- Given
  - prior distribution $\mathcal{N}_p(\beta_0, \Sigma)$ on $\beta$
  - the posterior distribution $\pi(\beta|y_1, \ldots, y_n, x_1, \ldots, x_n)$ is computed by

**Algorithm A.5 –Probit posterior distribution–**

1. `Simulate`

$$y_i^* \sim \begin{cases} \mathcal{N}_+(x_i^t\beta, 1, 0) & \texttt{if } y_i = 1, \\ \mathcal{N}_-(x_i^t\beta, 1, 0) & \texttt{if } y_i = 0, \end{cases} \qquad (i = 1, \ldots, n)$$

2. `Simulate`

$$\beta \sim \mathcal{N}_p\left((\Sigma^{-1} + XX^t)^{-1}(\Sigma^{-1}\beta_0 + \sum_i y_i^* x_i), (\Sigma^{-1} + XX^t)^{-1}\right)$$

where

- $\mathcal{N}_+(\mu, \sigma^2, \underline{u})$ and $\mathcal{N}_-(\mu, \sigma^2, \overline{u})$ denote the normal distribution truncated on the left in $\underline{u}$, and the normal distribution truncated on the right in $\overline{u}$, respectively
- $X$ is the matrix whose columns are the $x_i$'s.

$\|$

- Incomplete observations arise in numerous settings.
  - A survey with multiple questions may include nonresponses to some personal questions;
  - A calibration experiment may lack observations for some values of the calibration parameters;
  - A pharmaceutical experiment on the aftereffects of a toxic product may skip some doses for a given patient.

- The analysis of such structures is complicated by the fact that the failure to observe is not always explained.

- If these missing observations are entirely due to chance, it follows that the incompletely observed data only play a role through their marginal distribution.

- However, these distributions are not always explicit and a natural approach leading to a Gibbs sampler algorithm is to replace the missing data by simulation.

## Example 9.1.2 –Non-ignorable non-response–

- Average incomes and numbers of responses/non-responses to a survey on the income by age, sex and marital status. (*Source:* Little and Rubin 1987.)

| | Men | | Women | |
|---|---|---|---|---|
| Age | Single | Married | Single | Married |
| $< 30$ | 20.0 | 21.0 | 16.0 | 16.0 |
| | 24/1 | 5/11 | 11/1 | 2/2 |
| $> 30$ | 30.0 | 36.0 | 18.0 | – |
| | 15/5 | 2/8 | 8/4 | 0/4 |

- The observations are grouped by average, and we assume an exponential shape for the individual data,

$$y^*_{a,s,m,i} \sim \mathcal{E}xp(\mu_{a,s,m}) \quad \text{with} \quad \mu_{a,s,m} = \mu_0 + \alpha_a + \beta_s + \gamma_m \ ,$$

  where
  - $1 \le i \le n_{a,s,m}$
  - $\alpha_a$ $(a = 1, 2)$ corresponds to *age* (junior/senior)
  - $\beta_s$ $(s = 1, 2)$ corresponds to *sex* (fem./male)
  - $\gamma_m$ $(m = 1, 2)$ corresponds to *family* (single/married)
- The model is unidentifiable, but that can be remedied by constraining $\alpha_1 = \beta_1 = \gamma_1 = 0$.

- A more difficult and important problem appears when nonresponse depends on the income, say in the shape of a logit model,

$$p_{a,s,m,i} = \frac{\exp\{w_0 + w_1 y^*_{a,s,m,i}\}}{1 + \exp\{w_0 + w_1 y^*_{a,s,m,i}\}},$$

  where

  $p_{a,s,m,i}$ denotes the probability of nonresponse and

  $(w_0, w_1)$ are the logit parameters.

- The likelihood of the complete model is

$$\prod_{\substack{a=1,2\\s=1,2\\m=1,2}} \prod_{i=1}^{n_{a,s,m}} \frac{\exp\{z^*_{a,s,m,i}(w_0 + w_1 y^*_{a,s,m,i})\}}{1 + \exp\{w_0 + w_1 y^*_{a,s,m,i}\}} (\mu_0+\alpha_a+\beta_s+\gamma_m)^{r_{a,s,m}}$$
$$\times \exp\left\{-r_{a,s,m}\overline{y}_{a,s,m}(\mu_0 + \alpha_a + \beta_s + \gamma_m)\right\}$$

  where

  - $z^*_{a,s,m,i}$ is the indicator of a missing observation
  - $n_{a,s,m}$ is the number of people by category
  - $r_{a,s,m}$ is the number of responses by category
  - $\overline{y}_{a,s,m}$ is the average of these responses by category

- The completion of the data then proceeds by simulating
  - The $y^*_{a,s,m,i}$'s from

$$\pi(y^*_{a,s,m,i}) \propto \exp(-y^*_{a,s,m,i}\,\mu_{a,s,m}) \frac{\exp\{z^*_{a,s,m,i}(w_0 + w_1 y^*_{a,s,m,i})\}}{1 + \exp\{w_0 + w_1 y^*_{a,s,m,i}\}},$$

    which requires a Metropolis–Hastings step.
  - The parameters are simulated from

$$\prod_{\substack{a=1,2\\ s=1,2\\ m=1,2}} (\mu_0 + \alpha_a + \beta_s + \gamma_m)^{r_{a,s,m}} \times \exp\left\{-r_{a,s,m}\overline{y}_{a,s,m}(\mu_0 + \alpha_a + \beta_s + \gamma_m)\right\}$$

    for $\mu_0, \alpha_2, \beta_2, \gamma_2$, possibly using a gamma instrumental distribution.
  - And $(w_0, w_1)$ from

$$\prod_{\substack{a=1,2\\ s=1,2\\ m=1,2}} \prod_{i=1}^{n_{a,s,m}} \frac{\exp\{z^*_{a,s,m,i}(w_0 + w_1 y^*_{a,s,m,i})\}}{1 + \exp\{w_0 + w_1 y^*_{a,s,m,i}\}}$$

    which corresponds to a logit model.

$\|$

### 9.2 Finite mixtures of distributions

- *Mixtures of distributions*
$$\tilde{f}(x) = \sum_{j=1}^{k} p_j \ f(x|\xi_j) \ ,$$
where $p_1+\ldots+p_k = 1$, are useful in practical modeling.
- They can be challenging from an inferential point of view, that is, when estimating the parameters $p_j$ and $\xi_j$.
- The likelihood is quite difficult to work with, being of the form
$$L(p,\xi|x_1,\ldots,x_n) \propto \prod_{i=1}^{n} \left\{ \sum_{j=1}^{k} p_j \ f(x_i|\xi_j) \right\} \ ,$$
containing $k^n$ terms.
- A solution is to take advantage of the missing data structure, and associate with every observation $x_i$ an indicator variable $z_i \in \{1,\ldots,k\}$ that indicates which component of the mixture $x_i$ comes from. The demarginalization (or *completion*) of the mixture model is then
$$z_i \sim \mathcal{M}_k(1; p_1,\ldots,p_k), \qquad x_i|z_i \sim f(x|\xi_{z_i}) \ .$$
- The likelihood of the completed model is
$$\ell(p,\xi|x_i^*,\ldots,x_i^*) \ \propto \ \prod_{i=1}^{n} \ p_{z_i} \ f(x_i|\xi_{z_i})$$

$$= \prod_{j=1}^{k} \prod_{i;z_i=j} p_j \; f(x_i|\xi_j)$$

• A Gibbs sampler is then

**Algorithm A.6 –Mixture simulation–**

1. `Simulate` $z_i \ (i = 1, \ldots, n)$ `from`
   $P(z_i = j) \propto p_j \ f(x_i|\xi_j) \qquad (j = 1, \ldots, k)$
   `and compute the statistics`
   $$n_j = \sum_{i=1}^{n} \ \mathbb{I}_{z_i=j} \ , \qquad n_j\overline{x}_j = \sum_{i=1}^{n} \ \mathbb{I}_{z_i=j} x_i \ .$$
2. `Generate` $(j = 1, \ldots, k)$
   $$\xi \ \sim \ \pi \left( \xi | \frac{\lambda_j \alpha_j + n_j \overline{x}_j}{\lambda_j + n_j}, \lambda_j + n_j \right),$$
   $$p \ \sim \ \mathcal{D}_k(\gamma_1 + n_1, \ldots, \gamma_k + n_k) \ .$$

**Example 9.2.1 – Normal mixtures** – In the case of a mixture of normal distributions,

$$\tilde{f}(x) = \sum_{j=1}^{k} \ p_j \ \frac{e^{-(x-\mu_j)^2/(2\tau_j^2)}}{\sqrt{2\pi} \ \tau_j} \ ,$$

the conjugate distribution on $(\mu_j, \tau_j)$ is

$$\mu_j|\tau_j \sim \mathcal{N}\left(\alpha_j, \tau_j^2/\lambda_j\right), \qquad \tau_j^2 \sim \mathcal{IG}\left(\frac{\lambda_j + 3}{2}, \frac{\beta_j}{2}\right)$$

and the two steps of the Gibbs sampler are as follows →

**Algorithm A.7 –Normal mixture–**

1. `Simulate` $(i = 1, \ldots, n)$

$$z_i \sim P(z_i = j) \propto p_j \ \exp\left\{-(x_i - \mu_j)^2/(2\tau_j^2)\right\} \ \tau_j^{-1}$$

`and compute the statistics` $(j = 1, \ldots, k)$

$$n_j = \sum_{i=1}^{n} \mathbb{I}_{z_i=j}, \quad n_j\overline{x}_j = \sum_{i=1}^{n} \mathbb{I}_{z_i=j} x_i, \quad s_j^2 = \sum_{i=1}^{n} \mathbb{I}_{z_i=j}(x_i - \overline{x}_j)^2 \ .$$

2. `Generate`

$$\begin{aligned} \mu_j|\tau_j &\sim \mathcal{N}\left(\frac{\lambda_j\alpha_j + n_j\overline{x}_j}{\lambda_j + n_j}, \frac{\tau_j^2}{\lambda_j + n_j}\right), \\ \tau_j^2 &\sim \mathcal{IG}\left(\frac{\lambda_j + n_j + 3}{2}, \frac{\beta_j + s_j^2}{2}\right), \\ p &\sim \mathcal{D}_k(\gamma_1 + n_1, \ldots, \gamma_k + n_k) \ . \end{aligned}$$

$\|$

**Example 9.2.2 –Stochastic Volatility–**

- Stochastic volatility models are popular in financial applications, especially in describing series with sudden changes in the magnitude of variation of the observed values.

- They use a latent linear process $(Y_t^*)$, called the *volatility*, to model the variance of the observables $Y_t$.

- Let $Y_0^* \sim \mathcal{N}(0, {\sigma^*}^2)$ and, for $t = 1, \ldots, T$, define
$$\begin{cases} Y_t^* = \varrho Y_{t-1}^* + \sigma^* \epsilon_{t-1}^* \ , \\ Y_t = e^{Y_t^*/2} \epsilon_t \ , \end{cases}$$
where $\epsilon_t$ and $\epsilon_t^* \sim \mathcal{N}(0,1)$.

- The observed likelihood $L(\varrho, \sigma^* | y_0, \ldots, y_T)$ is obtained by integrating the complete-data likelihood
$$L^c(\varrho, \sigma^* | y_0, \ldots, y_T, y_0^*, \ldots, y_T^*)$$
$$\propto \ \exp - \sum_{t=0}^{T} \left\{ y_t^2 e^{-y_t^*} + y_t^* \right\} / 2$$
$$\times (\sigma^*)^{-T+1} \exp - \left\{ (y_0^*)^2 + \sum_{t=1}^{T} (y_t^* - \varrho y_{t-1}^*)^2 \right\} / 2(\sigma^*)^2 \ .$$

- The figure shows a typical stochastic volatility behavior for $\sigma^* = 1$ and $\varrho = .9$.

- Likelihood and Bayesian inference on this model can be done with the EM algorithm or the Gibbs sampler

$\|$